Public Archives Canada | Archives publiques Canada

Canada

Machine Readable Archives BULLETIN

Vol. 4 – No. 1
Spring 1986

ISSN 0821-3658

Computerization of the 1871 Ontario Census

Early in 1982, the Ontario Genealogical Society (OGS) approved a proposal by its members Bruce Elliott of Ottawa and Laurena Storey of London that it undertake an index to the 1871 census of Ontario as a project to commemorate the organization's 25th anniversary in 1986. In the autumn of 1982, Bruce Elliott met with Professor John Clarke of the Geography Department of Carleton University and Harold Naugler and David Brown of the Machine Readable Archives Division (MRA) of the Public Archives to discuss making the project a joint endeavour between the Society and the MRA. The result was an agreement whereby the OGS agreed that its members would extract full information from the census (name, sex, age, birthplace, religion, origin and occupation) for every head of family and every individual bearing a surname different from the head (flagging the latter as "strays"), and that the MRA would convert the extracted information into machine-readable form. Each of the twenty-six OGS branches appointed a coordinator to oversee transcription in its own area and received listings and instructions from Bruce Elliott who served as the project's Provincial Coordinator. By the end of 1985, 98.5 per cent of the data extraction had been completed by more than 400 volunteers and computerized by the MRA, and the computerized data for more than half the province had been cross-checked against the microfilms of the original census schedules, again by OGS members.

The results of the project will be made available to the public in three stages. The OGS commenced publication in spring 1986 of a thirty-volume series of nominal indexes to the census on a county basis. The volumes for Halton-Peel and Huron were launched at the Society's annual Seminar in Windsor in May. These county volumes mark the first step toward completion of a province-wide A–Z index that will eventually be available from the Society on microform once verification of the data for the entire province has been completed. These indexes will greatly facilitate genealogical and biographical research by allowing an investigator to locate a given individual quickly, even if the place of residence is unknown. The indexes will therefore be of inestimable value to researchers and to descendants of Ontario families resident in the United States and the Canadian West, for whom the former residence in Ontario of their pioneer ancestors remains a mystery. It will also be of use to researchers in foreign countries searching for emigrants, and because of its comprehensiveness it will become the major source for determining the regionalization of surnames in the province. Its main use will probably be in locating migrants. Quite aside from its genealogical value, social scientists studying internal migration and social mobility will find they have gained a useful tool for tracing the later whereabouts of residents who left specific communities in the 1850s and 1860s.

The machine-readable data base created will become available through the Machine Readable Archives Division. When archived, it will be possible for researchers to request copies or extracts from the data base and to undertake statistical cross-tabulations for specialized research projects. For example, researchers will be able to prepare lists of families of African origin in the province, percentage tabulations of Irish or Scottish Catholics and Protestants in specific communities, lists of photographers in a given region, and so forth. Thus the project will result not only in the production of a comprehensive index to the population of Ontario just after Confederation but also in the creation of a computer-manipulable data base that will be of use to scholars and students in various disciplines.

Bruce S. Elliott

CULDAT

An article in Volume 2 – Number 4 of the *Bulletin* described the origins of the pilot project undertaken to develop a Canadian Union List of Machine Readable Data Files (CULDAT). The work on the development of the online inventory was contracted to the Social Science Computing Laboratory at the University of Western Ontario. The overall purpose was to develop organizational, technical and informational foundations for maintaining and disseminating a computerized inventory. Specific objectives involved: the establishment of a standard for describing MRDF for entry into the data base; the design and implementation of the pilot data base containing a partial inventory; and the definition of the organizational roles and mechanism to effect the routine and cost-effective flow of descriptive information from data archives and other organizations to the union list beyond the conclusion of the pilot.

The pilot project was carried out over a fourteen month period. In January 1985, a Committee of data archivists and data librarians established a list of elements that were to be used to describe the holdings of the institutions. These elements were taken from those defined in the MARC format for data files. A data dictionary was developed to aid participants in the entry of descriptive information. The Social Science Computing Laboratory was involved in six major activities: the creation of the pilot data base; the set-up of online access with Basis on the Lab's VAX11/785; the set-up of DATAPAC and standard dial-up communications; the conducting of an evaluation of the online system; a survey of potential contributors; and production of a hard copy reference document.

Contributors to the data base were from the university-based archives and libraries and included: Data Library, University of British Columbia; Institute of Social Research, York University; Data Resources Library, University of Western Ontario; Institute for Social and Economic Research, University of Manitoba. The MRA also contributed descriptive entries. In all, 753 records were entered into CULDAT. The evaluation of the data base was extended to more participants than those listed above and included both frequent users of online systems as well as infrequent users. Although a number of suggestions have been made on how to improve the online inventory, the general consensus was that the data base was very useful and should be continued.

It is of no surprise that the most crucial component of the data base was the description

cont'd on page 2

"CULDAT"

cont'd from page 1

of the data file. A number of difficulties were experienced with the lack of consistent terminology used and the detail of the description itself. A number of problems were encountered and are summarized in the following paragraphs. The resolution of these difficulties has formed the basis of the CULDAT work plan for 1986–87.

The choice of the data elements to be included in CULDAT was based on the fields of the MARC format for data files. A limited number of elements were chosen as it was felt by the Committee that the intention of the data base was to include only sufficient information to identify a unique data file, to aid researchers in selecting files of interest, and to locate archived copies of the file. The resulting CULDAT Data Element Dictionary contained the field names and a brief description. During the pilot project, it was noted that in some cases the data dictionary did not provide sufficient guidance to the archivist or librarian to describe the data files and presumed a knowledge of the MARC format and Anglo-American Cataloguing Rules II. This created some difficulty in mapping out the information received for input into CULDAT. The consequences of a weak data element dictionary are inconsistent presentation of the data that can make the descriptions difficult for the end user to interpret. Weak data descriptions yield inefficient indexes, which, in turn, require the user to anticipate all possible variations of a term in order to find all relevant records in the data base. Specific problems were found in the following data elements.

(1) *Investigators*—The differentiation between principal investigator and other investigators caused some difficulties for both the cataloguer and the user. The determination of principal investigator for a data file is difficult, if not impossible, at times. The separation of these fields requires searching two fields rather than one for the user wishing to browse the index. The distinction between investigator (personal) and investigator (corporate) was considered essential. The lack of authority control in the corporate investigator field was a problem that could be overcome through the use of Canadiana to control the use and spelling of names.

(2) *Producer; Generator; Distributor*—A tendency to repeat the same data in these fields was found. This may have been due to the inadequacy of the data dictionary. Abbreviations and acronyms were used. The adoption of an authority file for corporate names would apply to these fields as well.

(3) *File Size; Number of Cases*—Some difficulty was experienced in the data provided in this field. Again, this is due to the lack of guidance in the data dictionary.

(4) *Access Restrictions*—As all institutions have their own access regulations, it was felt that this field should only be completed when the distributing organization has contributed the record.

(5) *Abstract*—Information contained in this field was found at times to repeat information found in other fields. The vocabulary used varied widely, which made control of the field extremely difficult. The types of variables used in a data file is vital information for the prospective user. In order to provide improved access to this field, it would be preferable to separate the abstract from the variable list. Variables could then be indexed as phrases rather than individual words, and the abstract could be left unindexed. Such a change would significantly reduce the indexing overhead and improve the quality of the printed keyword index by using variable names instead of individual words. The online system could continue to index variables as individual words as well as expressions.

(6) *Geographic Coverage*—The pattern adopted by the pilot was as follows: site, city, region, territory, province, state, country (qualifier), continent. The pattern worked well in most cases and ensured that the user interested in data about a particular province could retrieve information on a file that covered only a city in that province. The only records that do not conform to this pattern are physical data where orbital coordinates are submitted.

(7) *Chronological Coverage*—The format of the dates recorded in this field were inconsistent, rendering the retrieval of data ineffective. The data dictionary should prescribe one acceptable format and all dates would be converted. The standard format will provide the possibility of performing systematic retrieval on time periods by scanning the text, even though every unit of time within a range is not actually recorded in the field.

The difficulties that have been encountered will provide valuable information to improve the quality and guidance required for the data dictionary. The second version should improve the consistency of the descriptive entries. The contributions made by the data archives and libraries were extremely useful in building the pilot data base and being able to identify specific needs to improve the data dictionary.

User Evaluation and Potential Contributors

The original project design called for online testing and evaluation of the pilot CULDAT data base by project participants and preparing a list and contacting potential contributing organizations in order to learn about their holdings and interest in submitting entries into CULDAT in the future. Three important additions were made to enhance the project. The establishment of a DATAPAC Service reduced usage costs and significantly improved convenience to remote users. In addition, the survey of contributors was expanded to include questions on evaluation as they were potential users as well. The third activity was to include three local University of Western Ontario groups (students in the School of Library and Information Science, the University's reference librarians, and social science researchers who use the Lab's support services). These additions increased the use of CULDAT during the pilot phase.

The evaluation of the data base was very favourable and many respondents expected to benefit from the availability of CULDAT in the future. Considerable information from prospective contributors and users was acquired. This information and experience provides a sound foundation for the design and planning of the next stages in the development of CULDAT. The MRA would like to thank all of those who contributed their time through the preparation of descriptions of their holdings, the testings and evaluation of the data base, and their response as potential contributors.

This article, by Sue Gavrel, is an abridged version of the Final Report, "Pilot Project for the Development of a Canadian Union List of Machine Readable Data Files (CULDAT)," prepared by Edward H. Hanis, Social Science Computing Laboratory, University of Western Ontario for the Machine Readable Archives, Public Archives of Canada.

Anyone wishing information about the *Bulletin* may write to: Public Archives of Canada, Machine Readable Archives Division, Documentation and Public Service Section, 395 Wellington Street, Ottawa, Ontario K1A 0N3, or phone (613) 993-7772.

Il n'est pas étonnant que l'élément le plus important de la base de données ait été la description du fichier. Le manque d'uniformité dans la terminologie et le peu de détails descriptifs ont occasionné quelques difficultés qui sont décrites dans les paragraphes qui suivent. La résolution de ces problèmes servira de base au plan de travail du CULDAT en 1986–1987.

Le choix des éléments de données s'est fait à partir des éléments descriptifs du format MARC pour les fichiers de données. Un petit nombre d'éléments ont été choisis, car le comité estimait que la base de données ne devait comprendre que l'information nécessaire pour identifier un fichier et pour aider les chercheurs à choisir des fichiers pertinents et à localiser les copies archivistiques du fichier. Le dictionnaire des éléments de données CULDAT contenait les titres des zones et une brève description de celles-ci. Lors du projet-pilote, on a constaté que dans certains cas le dictionnaire ne fournissait pas suffisamment de directives à l'archiviste ou au bibliothécaire pour lui permettre de décrire les fichiers de données, et présupposait une connaissance du format MARC et de la deuxième édition des Règles de catalogage anglo-américaines. La disposition des données destinées au CULDAT s'en est trouvée compliquée du même coup. Il est difficile de présenter les données de façon cohérente si le dictionnaire est inadéquat au départ, ce qui fait que l'utilisateur final peut difficilement interpréter les descriptions fournies. Les descriptions de piètre qualité donnent de mauvais index qui, à leur tour, obligent l'usager à prévoir toutes les variations possibles d'un terme de façon à trouver toutes les notices pertinentes dans la base de données. Des problèmes se sont posés dans les éléments de données suivants.

1) *Enquêteurs*—La distinction entre l'enquêteur principal et les autres enquêteurs pose des problèmes tant au catalogueur qu'à l'utilisateur. Il est parfois difficile, voire impossible, d'identifier le principal enquêteur d'un fichier de données. La séparation de ces zones oblige l'utilisateur à consulter deux zones plutôt qu'une lorsqu'il désire parcourir l'index. La distinction entre enquêteur (individuel) et enquêteur (collectif) est considérée comme essentielle. L'absence de contrôle d'autorité dans la zone enquêteur collectif est un problème qui pourrait être réglé par l'utilisation de *Canadiana* pour vérifier l'emploi et l'orthographe des noms.
2) *Producteur, Distributeur*—On a tendance à répéter les mêmes données dans ces zones et ce, peut-être à cause d'une faille dans le dictionnaire de données. Des abréviations et des sigles ont été utilisés. L'adoption d'un fichier d'autorité pour les noms de collectivités s'appliquerait également à ces zones.
3) *Importance du fichier, Nombre de cas*—Les données de cette zone ont posé un problème et, encore là, à cause de l'absence de directives dans le dictionnaire de données.
4) *Restrictions concernant l'accès*—Comme tous les établissements ont leurs propres règlements en la matière, cette zone devrait être remplie seulement si c'est l'organisme diffuseur qui a fourni la notice.
5) *Résumé analytique*—Les renseignements contenus dans cette zone répètent souvent ceux consignés dans d'autres zones. Les termes utilisés varient considérablement, ce qui rend le contrôle de cette zone très difficile. Les types de variables utilisés dans un fichier de données sont des renseignements essentiels pour l'utilisateur éventuel. On aurait donc tout intérêt à séparer le résumé analytique de la liste de variables. Les variables pourraient ensuite être indexées à titre d'expressions plutôt que de mots et le résumé analytique resterait sans index. Ce changement réduirait considérablement les frais d'indexage en plus d'améliorer la qualité de l'index imprimé des mots-clés parce que des variables seraient substituées à des mots. Le système en direct continuerait à indexer les variables comme mot ou comme expression.
6) *Champ géographique*—Le mode suivant a été adopté dans le cadre du projet-pilote : emplacement, ville, région, territoire, province, État, pays (qualificatif), continent. Cette façon de procéder a bien fonctionné dans la plupart des cas et a permis à l'utilisateur qui ne voulait que des données sur une province particulière d'être effectivement renseigné sur un fichier portant uniquement sur une ville de cette province. Les seules notices qui ne se conforment pas à ce procédé sont les données matérielles pour lesquelles des coordonnées sphériques entrent en ligne de compte.
7) *Champ chronologique*—La structure d'enregistrement des dates dans cette zone est inconséquente, ce qui rend la recherche des données inefficace. Le dictionnaire de données devrait prescrire une forme acceptable et toutes les dates y seraient converties. La forme standard permettra des recherches systématiques sur la durée des sondages en scrutant le texte, même si chaque unité de temps à l'intérieur d'une série n'est pas réellement enregistrée dans une zone.

Toutes ces difficultés qu'il faudra surmonter nous permettront d'améliorer la qualité du dictionnaire de données. Les notices descriptives de la deuxième version seront plus cohérentes. La contribution des informathèques et des archives de données informatiques a été très utile puisqu'elle a permis de monter la base de données pilote et de préciser les points à améliorer dans le dictionnaire de données.

Évaluation par les utilisateurs et participants éventuels

Les concepteurs du projet ont prévu que les participants feraient l'essai et l'évaluation de la base de données CULDAT en direct et qu'une liste des organismes intéressés serait dressée pour qu'on puisse les contacter et se renseigner sur leurs fonds. Trois éléments se sont ajoutés au projet dans le but de l'améliorer. La création d'un service DATAPAC a réduit les coûts d'utilisation et facilité la tâche des usagers éloignés. De plus, l'enquête menée auprès des participants a été élargie pour y inclure des questions sur l'évaluation parce qu'ils sont aussi des utilisateurs éventuels. La troisième activité a porté sur la participation de trois groupes affiliés à l'Université Western Ontario (les étudiants inscrits à l'École de bibliothéconomie et des sciences de l'information, les bibliothécaires à la référence et les spécialistes des sciences sociales qui ont recours aux services de soutien du laboratoire). Grâce à ces ajouts le CULDAT a été utilisé plus souvent au cours du projet-pilote.

L'évaluation de la base de données s'est avérée très positive et de nombreux répondants s'attendent à profiter du CULDAT à l'avenir. Les participants et utilisateurs éventuels ont fourni quantité d'informations qui serviront de base à la conception et à la planification des prochaines étapes du projet. La DAO remercie tous ceux qui ont pris le temps de décrire leurs fonds, de faire l'essai de la base de données et de l'évaluer et qui ont manifesté l'intérêt d'y participer éventuellement.

Traduction d'une version abrégée par Sue Gavrel du rapport final « Pilot Project for the Development of a Canadian Union List of Machine Readable Data Files (CULDAT) », qu'a préparé Edward H. Hanis, du Social Science Computing Laboratory de l'Université Western Ontario pour les Archives ordinolingues des Archives publiques du Canada.

Pour obtenir des renseignements au sujet du *Bulletin*, il suffit d'écrire à l'adresse suivante : Archives publiques du Canada, Division des archives ordinolingues, Section de la documentation et du service au public, 395, rue Wellington, Ottawa (Ontario), K1A 0N3, ou de téléphoner au (613) 993-7772.

Archives publiques Canada
Public Archives Canada

Canada

Vol. 4 – n° 1
Printemps 1986

Archives ordinolingues BULLETIN

ISSN 0821-3658

Automatisation des données du recensement de 1871 de l'Ontario

Au début de 1982, l'Ontario Genealogical Society (OGS) décidait sur une proposition de deux de ses membres, Bruce Elliott d'Ottawa et Laurena Storey de London, d'établir un index pour le recensement de 1871, en vue de souligner le 25e anniversaire de la Société en 1986. À l'automne 1982, Bruce Elliott s'est entretenu avec le professeur John Clarke du Département de géographie de l'Université Carleton ainsi qu'avec Harold Naugler et David Brown de la Division des archives ordinolingues (DAO) des Archives publiques pour faire du projet d'indexage une entreprise commune de l'OGS et de la DAO. L'OGS a donc convenu d'extraire tous les renseignements (nom, sexe, âge, lieu de naissance, religion, origine et profession) sur chaque chef de famille et chaque personne portant un patronyme qui différait de celui du chef de famille (ces personnes étant considérées comme « errantes »). Pour sa part, la DAO a convenu de convertir les données extraites sur supports ordinolingues. Chacune des 26 sections de l'OGS a nommé un coordonnateur pour superviser la transcription des données dans sa région et a reçu des listes et des directives de Bruce Elliott qui a fait fonction de coordonnateur provincial. Au 31 décembre dernier, 98,5 % des données avaient été extraites par plus de 400 bénévoles puis informatisées par la DAO, et plus de la moitié des données de la province avaient été contre-vérifiées par des membres de l'OGS qui les avaient comparées aux bordereaux du recensement original.

Les résultats du projet seront diffusés en trois étapes. Au cours du printemps, l'OGS a commencé à publier une série d'index nominatifs en 30 volumes, par comté. Les volumes pour les comtés de Halton-Peel et de Huron ont été présentés lors du colloque annuel de la Société qui s'est tenu à Windsor en mai. Ces volumes sont les premiers d'un index alphabétique de tous les comtés de la province que l'OGS produira sur microforme dès que toutes les données auront été vérifiées. Cet index facilitera considérablement les recherches généalogiques et biographiques en permettant aux enquêteurs de localiser telle ou telle personne rapidement, même si son lieu de résidence est inconnu. Il sera donc d'une valeur inestimable pour les chercheurs et les descendants des familles ontariennes établis aux États-Unis et dans l'Ouest canadien, qui voudraient connaître le lieu de résidence de leurs ancêtres en Ontario. Par ailleurs, les chercheurs dans les pays étrangers pourront le consulter pour se renseigner sur l'identité des émigrants, et en raison de sa complétude l'index constituera la principale source documentaire sur la répartition géographique des noms de famille dans la province. Il est probable que l'index servira avant tout à localiser des migrants. Les spécialistes des sciences sociales qui étudient les migrations internes et la mobilité sociale pourront s'en servir pour retracer les personnes qui ont quitté des régions particulières dans les années 1850 et 1860.

La base de données ainsi créée sera accessible par l'entremise de la Division des archives ordinolingues. Une fois les données traitées à des fins archivistiques, les chercheurs pourront demander des copies ou extraits de cette base de données pour dresser des tableaux statistiques à double entrée dans le cadre de projets spécialisés. Ainsi, les chercheurs pourront dresser toutes sortes de listes, p. ex. celles des familles d'origine africaine dans la province ou des photographes qui habitaient une région donnée, ou pourront déterminer le pourcentage de catholiques et de protestants irlandais ou écossais dans certaines localités. Le projet n'entraînera pas seulement la production d'un index exhaustif sur la population de l'Ontario après la Confédération, mais aussi la création d'une base de données informatisées dont pourront tirer profit une multitude de spécialistes et d'étudiants.

Bruce S. Elliott

CULDAT

Dans un *Bulletin* antérieur (vol. 2, n° 4), on décrivait les origines du projet-pilote visant à dresser un Catalogue collectif canadien de fichiers de données ordinolingues (CULDAT). Un contrat a été passé avec le Social Science Computing Laboratory de l'Université Western Ontario en vue d'établir une base organisationnelle, technique et informative pour la tenue et la diffusion de l'inventaire informatisé. Il y avait certains objectifs précis à atteindre, notamment le choix d'une norme descriptive pour l'introduction des fichiers de données ordinolingues, la conception et la mise en place de la base de données pilote contenant un inventaire partiel, et la définition des rôles et mécanismes organisationnels liés à une circulation courante et rentable des données descriptives en provenance d'archives de données et d'organismes analogues, même après la conclusion du projet-pilote.

Le projet-pilote dura quatorze mois. En janvier 1985 un comité d'archivistes et de bibliothécaires-statisticiens dressait une liste d'éléments qui devaient servir à décrire les collections des organismes visés. Ces éléments émanaient du format MARC utilisé pour les fichiers de données. Un dictionnaire de données a été élaboré pour aider les participants à introduire leurs données descriptives. Le Social Science Computing Laboratory a participé à six grandes activités : la création d'une base de données pilote, la constitution de l'accès en direct avec le Basis à l'aide de l'ordinateur VAX11/785, l'organisation du réseau DATAPAC et de communications commutées, l'évaluation du système en direct, la tenue d'une enquête auprès des participants éventuels et la production d'un manuel de référence.

La base de données a été alimentée par des bibliothèques et des établissements d'archives affiliés à des universités, y compris : la Data Library de l'Université de la Colombie-Britannique, l'Institute of Social Research de l'Université York, la Data Resources Library de l'Université Western Ontario et l'Institute for Social and Economic Research de l'Université du Manitoba. La DAO a également fourni des données descriptives et, en tout, 753 notices ont été introduites dans le CULDAT. L'évaluation de la base de données a été confiée à de nombreux autres participants, aussi bien ceux qui se servent fréquemment des systèmes en direct que ceux qui les utilisent moins souvent. Plusieurs suggestions ont été formulées quant à la façon d'améliorer l'inventaire en direct mais la plupart des participants ont toutefois estimé que la base de données était très utile et ne devait pas être abolie.

suite à la page 2